

ENT-ICIPATE

Samesun Singh
Politecnico di Torino
Turin, Italy

Alessandro Carrabs
Politecnico di Torino
Turin, Italy

Xiao Quan Ji
Politecnico di Torino
Turin, Italy

Abstract—Post-operative complications in Ear-Nose-Throat (ENT) oncology surgery, such as nosocomial infections and pharyngo-/oro-cutaneous fistulas, represent rare but high-impact clinical events. Early identification of high-risk patients is essential to improve post-surgical safety and optimize healthcare resources.

This project investigates the use of machine learning models to predict post-surgical complications using heterogeneous clinical data from 574 ENT oncology patients. A robust preprocessing pipeline was designed to address missing values, feature redundancy, and severe class imbalance. The models evaluated for this task include Logistic Regression, Random Forest, XGBoost, and CatBoost.

Model performance was assessed using imbalance-aware metrics such as ROC-AUC, precision, recall, F1-score, and PR-AUC. A safety-first threshold optimization strategy was adopted to prioritize recall and minimize missed complications. Results show that ensemble-based models achieve strong predictive performance while maintaining clinical interpretability through explainability techniques, with recall values reaching up to 0.86 for pharyngo-/oro-cutaneous fistula and 0.85 for nosocomial infection at optimized operating points.

The proposed framework demonstrates the potential of leveraging machine learning algorithms to predict clinical complications. The full implementation of the project is available at: <https://github.com/adsp-polito/2026-P6-ENT-icipate>

1 Introduction

Post-operative complications in ENT oncology surgery, such as pharyngo-/oro-cutaneous fistulas and nosocomial infections, pose a serious threat to patient safety, prolong hospitalization, and increase healthcare costs. Traditional risk assessment methods rely on isolated clinical indicators and physician experience, which may not fully capture complex interactions among patient, tumor, and surgical factors. Machine learning offers a promising approach to identify high-risk patients by learning these complex patterns directly from data. However, predicting rare complications presents challenges, including severe class imbalance, limited model interpretability, and the need for clinically meaningful decision thresholds. In this work, we propose an interpretable machine learning framework that addresses these challenges through robust preprocessing, ensemble-based models, threshold optimization guided by clinical priorities, and explainability techniques, such as SHAP, to support transparent and clinically informed decision-making. This framework aims not only to provide reliable predictions of post-operative complications, but also to offer interpretable insights

that help clinicians anticipate risks and deliver more personalized care.

2 Related Work

Machine learning has been increasingly applied to support clinical research and outcome prediction in surgical and oncological settings. Data-driven approaches can model complex and non-linear relationships among patients, tumors, and procedure-related variables, providing a complement to traditional statistical risk assessment methods [9], [10].

In the context of otolaryngology, Rajan et al. [9] provide a broad overview of current trends and applications of artificial intelligence, highlighting both diagnostic and predictive use cases. Park et al. [10] systematically review the use of machine learning models in head and neck cancer surgery, emphasizing their potential for predicting surgical margins, postoperative complications, and the need for salvage procedures. Their review highlights that machine learning may be particularly useful in high-risk or low-prevalence scenarios, where rule-based linear models, like Linear Regression [5] could fail to capture complex clinical patterns. Van Dung et al. [8] provide a scoping review on the use of AI for surgical training and skill assessment in otolaryngology, illustrating broader applications of AI in the field, though not directly related to clinical outcome prediction.

Tree-based ensemble methods are widely adopted for clinical outcome prediction. Breiman [1] introduced Random Forests, effective in handling heterogeneous clinical variables and non-linear interactions. Friedman [2] proposed Gradient Boosting, a sequential learning strategy that iteratively corrects model errors. Implementations such as XGBoost by Chen and Guestrin [3] and CatBoost by Prokhorenkova et al. [4] enhance scalability and performance, making them suitable for structured clinical datasets.

Severe class imbalance represents a key challenge in medical prediction tasks, as adverse outcomes such as post-operative complications are relatively rare. Saito and Rehmsmeier [6] showed that precision-recall curves provide a more informative evaluation than ROC curves in imbalanced settings. Several clinical studies have documented that standard evaluation metrics such as accuracy or ROC-AUC are often reported without explicitly adopting imbalance-aware assessment or exploring clinically guided threshold selection strategies [11], [12].

Interpretability is another important requirement for clinical adoption. Lundberg and Lee [7] introduced SHAP, a unified framework for explaining individual predictions and global model behavior, which has become widely used for enhancing transparency and trust. Existing studies often apply interpretability methods independently, without integrating them with approaches for handling class imbalance or rare-event prediction.

Building upon this literature, this work combines robust preprocessing, imbalance-aware evaluation metrics, clinically guided threshold optimization, and interpretable machine learning models within a unified framework for predicting post-operative complications in ENT oncology surgery.

3 Method

3.1 Problem Formulation

We address the prediction of rare post-operative complications in ENT oncology patients, which remain relatively uncommon in our dataset as shown in the class distribution after removing patients with missing targets (Figure 1). This is formulated as a binary classification task. Each patient is represented by a feature vector $\mathbf{x} \in \mathbb{R}^d$, including demographic, clinical, tumor-related, and surgical variables. The target variable $y \in \{0, 1\}$ indicates whether a complication, such as pharyngo-/oro-cutaneous fistula or nosocomial infection, occurred.

Due to the high clinical cost of missed complications (false negatives), the modeling strategy emphasizes sensitivity to rare events.

3.2 Method Overview

The proposed framework consists of four main stages: data preprocessing, model training, threshold optimization, and interpretability analysis. After preprocessing, multiple machine learning models are trained and evaluated, thresholds are tuned to prioritize clinically relevant performance, and interpretability techniques are applied to identify key predictors of complications.

3.3 Data Preprocessing

The dataset includes heterogeneous clinical variables collected from multiple sources, resulting in inconsistent encodings, missing values, and potential redundancy. Categorical variables and target labels were standardized, and tumor staging information was harmonized into a unified ordinal format. Records lacking essential information were removed, while missing values in clinically meaningful variables were encoded as -1 to preserve informative missingness. Highly correlated or redundant features representing the same clinical concept were consolidated to reduce noise and improve model stability.

All preprocessing steps were applied consistently across training, validation, and test sets to prevent information leakage.

3.4 Handling Class Imbalance

Both target outcomes are rare clinical events, leading to a strong imbalance between positive and negative classes, as shown in Figure 1. Without proper adjustments, models tend to favor the majority class, reducing sensitivity to clinically important complications.

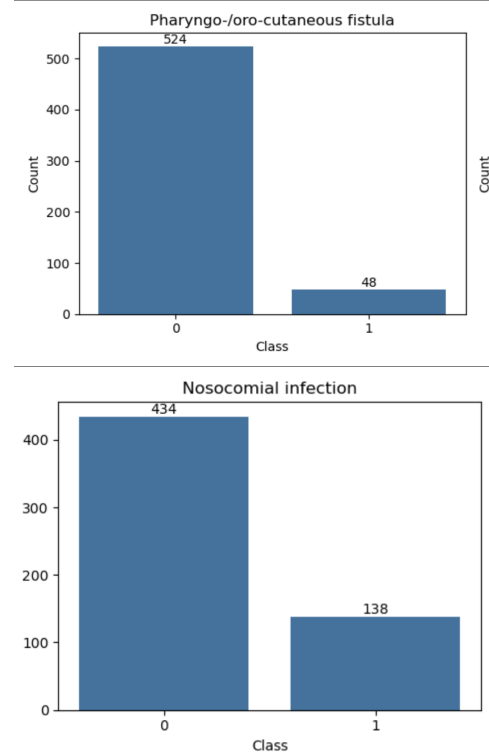


Fig. 1: Class distribution for post-surgical complications (0 = no complication, 1 = complication) after removal of patients with missing target values.

To mitigate this effect, the framework employs stratified data splitting to maintain class proportions across training, validation, and test sets, uses imbalance-aware learning algorithms, focuses evaluation on minority-class performance metrics, and optimizes classification thresholds with a safety-first perspective.

3.5 Machine Learning Models and Comparison

Several machine learning models were trained and compared to identify the best-performing approach. Class imbalance was explicitly addressed by setting the *class_weight* hyperparameter to 'balanced' for all models.

- Logistic Regression [5] was used as a baseline due to its simplicity and interpretability. The model estimates the log-odds of complication occurrence as a linear combination of features.

While limited in capturing non-linear interactions, logistic regression provides a reference point for evaluating the added value of more complex models.

- Random Forest [1] is an ensemble of decision trees trained on bootstrap samples, with feature subsampling at each split. This model captures non-linear interactions and is robust to noise.
- XGBoost [3] implement additive tree-based models where each new tree corrects the errors of the previous ensemble. XGBoost further incorporates regularization and advanced handling of missing values. These methods are particularly suited to heterogeneous clinical data and class imbalance, as the boosting procedure can focus on misclassified minority-class examples.
- CatBoost [4] is a gradient boosting algorithm that natively handles categorical variables without extensive preprocessing and reduces bias introduced by feature encoding. Its robustness to imbalanced and heterogeneous data makes it ideal for clinical datasets.

3.6 Hyperparameter Tuning and Threshold Optimization

Hyperparameters were optimized using stratified 5-fold cross-validation on the training set, targeting the average recall score.

Classification thresholds were subsequently tuned to adjust the classification threshold for converting predicted probabilities $\hat{p}_i = f(x_i)$ into binary predictions:

$$\hat{y}_i = \hat{y}_i(\tau) = \begin{cases} 1 & \text{if } \hat{p}_i \geq \tau \\ 0 & \text{if } \hat{p}_i < \tau \end{cases}$$

The optimal threshold τ^* was selected for each model by maximizing the F1-score for the positive class.

This procedure improves sensitivity to rare complications while controlling false positives, aligning the models with a safety-first clinical strategy. A *safety-first* operating point was adopted, reflecting the higher clinical cost of false negatives (missed complications) relative to false positives (additional alerts). The decision threshold was tuned on the validation set by maximizing the positive-class F1-score, rather than using the default threshold of 0.5, to obtain a clinically meaningful balance between sensitivity and precision under severe class imbalance.

4 Experiment

4.1 Dataset Description

The dataset consists of clinical records from 574 patients who underwent ENT oncology surgery at a single tertiary-care center. Each patient is described by 64 features derived from routinely collected and anonymized clinical data.

The variables are grouped into the following categories:

- **Patient demographics and lifestyle:** age, sex, anthropometric measures, smoking and alcohol-related variables.
- **Comorbidities:** including cardiovascular disease, diabetes, respiratory conditions, and other chronic illnesses relevant to surgical risk.
- **Tumor and staging information:** TNM staging variables originally encoded in heterogeneous and partially unstructured formats.
- **Surgical and operative characteristics:** type of surgical procedure, reconstruction techniques, flap usage, and operative complexity indicators.

Two binary target variables are considered: pharyngo-/oro-cutaneous fistula and nosocomial infection.

4.2 Experimental Configuration and Model Evaluation

The dataset was split into 70% training, 15% validation, and 15% test sets using stratified sampling to preserve the proportion of positive cases. All preprocessing steps described in Section 3.3 were applied consistently across splits.

Hyperparameters for each model were optimized using 5-fold stratified cross-validation on the training set, targeting the average recall score to account for class imbalance. Classification thresholds were subsequently tuned on the validation set to maximize the F1-score of the positive class,¹ reflecting a safety-first clinical approach.

Models were then compared on the validation set using clinically relevant metrics. The best-performing model was retrained on the combined training and validation sets and evaluated on the hold-out test set to assess its final performance.

Additionally, SHAP values were computed to quantify the contribution of each feature to individual predictions, providing global insights into feature importance and patient-specific explanations to support clinical decision-making [7].

Table III reports hold-out test set performance for the selected models, evaluated at the optimized thresholds.

4.3 Evaluation Metrics

Model performance was assessed using metrics sensitive to rare events, focusing on the positive class (patients experiencing complications). Overall accuracy was not considered due to strong class imbalance:

- **Recall (sensitivity):** measures the proportion of true complications correctly identified, emphasizing the minimization of false negatives.
- **Precision:** quantifies the proportion of predicted complications that are true, controlling false positives.

¹Candidate thresholds in the range [0.05, 0.9] with a step size of 0.01 were evaluated.

Model	F1-pos	Recall-pos	Precision-pos	PR-AUC	ROC-AUC
<i>Pharyngo-/Oro-Cutaneous Fistula</i>					
Random Forest	0.429	0.429	0.429	0.474	0.890
XGBoost	0.444	0.857	0.300	0.449	0.897
CatBoost	0.308	0.286	0.333	0.372	0.828
Logistic Regression	0.279	0.857	0.167	0.288	0.846
<i>Nosocomial Infection</i>					
Random Forest	0.619	0.850	0.486	0.625	0.862
XGBoost	0.596	0.700	0.519	0.608	0.838
CatBoost	0.609	0.700	0.538	0.671	0.866
Logistic Regression	0.604	0.800	0.485	0.620	0.845

TABLE I: Validation performance using the default threshold of 0.5.

Model	Threshold	F1-pos	Recall-pos	Precision-pos	PR-AUC	ROC-AUC
<i>Pharyngo-/Oro-Cutaneous Fistula</i>						
Random Forest	0.418	0.556	0.714	0.455	0.474	0.890
XGBoost	0.683	0.500	0.714	0.385	0.449	0.897
CatBoost	0.116	0.444	0.571	0.364	0.372	0.828
Logistic Regression	0.579	0.500	0.714	0.385	0.288	0.846
<i>Nosocomial Infection</i>						
Random Forest	0.607	0.636	0.700	0.583	0.625	0.862
XGBoost	0.418	0.615	0.800	0.500	0.608	0.838
CatBoost	0.381	0.653	0.800	0.552	0.671	0.866
Logistic Regression	0.484	0.630	0.850	0.500	0.620	0.845

TABLE II: Validation performance after threshold optimization. Thresholds were selected to maximize positive-class F1.

Target	Model	F1-pos	Recall-pos	Precision-pos	PR-AUC	ROC-AUC
Pharyngo-/Oro-Cutaneous Fistula	Random Forest	0.522	0.857	0.375	0.331	0.874
Nosocomial Infection	CatBoost	0.653	0.762	0.571	0.603	0.804

TABLE III: Test set performance of selected models using optimized thresholds.

- **F1-score:** balances recall and precision, providing a single indicator of model performance for the positive class.
- **PR-AUC:** area under the precision-recall curve, used as the primary evaluation metric due to class imbalance.
- **ROC-AUC:** area under the receiver operating characteristic curve, reported for completeness.

5 Results

5.1 Quantitative Results

We evaluated all trained models on the validation set, focusing on metrics related to the positive class due to the strong class imbalance in both targets. Performance was assessed using Precision, F1-score, PR-AUC, and ROC-AUC, as defined in Section 4.3

Tables I and II summarize the validation results using the default threshold of 0.5 and after threshold optimization, respectively, with the best values highlighted in **bold**.

Threshold optimization substantially improved detection of rare complications. In some cases, it also changed the ranking of the models in terms of F1-score, highlighting the importance of selecting thresholds appropriate for each outcome. For example, for pharyngo-/oro-cutaneous fistula, Random Forest im-

proved its F1-score from 0.429 at the default threshold to 0.556 after optimization, outperforming the other models. Similarly, for nosocomial infection, CatBoost achieved the highest F1-score after threshold tuning, surpassing models that performed better at the default threshold.

Overall, PR-AUC and ROC-AUC values remained high across thresholds and datasets, confirming good discrimination and reliable generalization.

5.2 Qualitative Results

Random Forest was particularly suitable for pharyngo-/oro-cutaneous fistulas, which are extremely rare. Its ensemble nature allows it to capture rare positive cases efficiently, supporting early detection in a clinical setting. CatBoost performed best for nosocomial infections, which are rare but more frequent than fistulas. Its ability to model complex feature interactions enables a better balance between sensitivity and precision, making it suitable for real-world monitoring of these complications. Figure 2 illustrates PR and ROC curves for the selected models on the test set, highlighting the practical impact of threshold selection.

If the models were deployed using the default threshold of 0.5 (orange markers), they would suffer from excessive conservatism, failing to identify the majority of patients who actually develop complica-

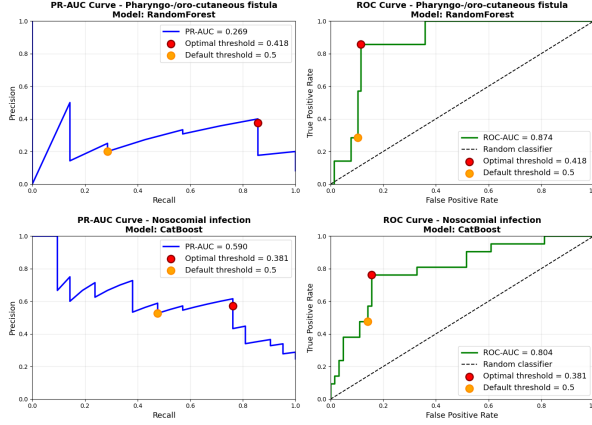


Fig. 2: ROC and PR curves on test set for the selected models. Top: pharyngo-/oro-cutaneous fistula. Bottom: nosocomial infection. The red points indicate the optimized thresholds, while the orange points correspond to the default threshold of 0.5.

tions. Instead, by utilizing the optimized thresholds (red markers), the models successfully navigate the trade-off between sensitivity and precision. In the PR space, this recalibration allows the system to capture rare events that would otherwise be overlooked, while in the ROC space, the models “climb” vertically to maximize the True Positive Rate. This shift ensures a significant gain in complication detection with a manageable increase in false alarms.

6 Model Interpretation

To analyze model behavior on the test set, we computed SHAP values [7] for each feature, which quantify its contribution to the predicted probability. Positive SHAP values indicate an increase in predicted risk, while negative values indicate a reduction relative to the model’s expected output.

All features were normalized to enable comparison across variables with different scales. Figure 3 displays the SHAP beeswarm plots for the two selected models. Each point corresponds to a patient, with horizontal position reflecting the impact on the predicted probability and color representing the normalized feature value. Features are ranked by mean absolute SHAP value to indicate overall importance.

For *pharyngo-/oro-cutaneous fistula*, the most influential features included tumor stage variables and specific surgical factors, showing relatively consistent effects across patients.

In contrast, for *nosocomial infection*, features such as age and comorbidity-related variables exhibited more heterogeneous contributions, suggesting interactions with other patient characteristics. These visualizations provide both a global overview of feature importance and insight into patient-level variability, supporting interpretation of model predictions in a clinical context.

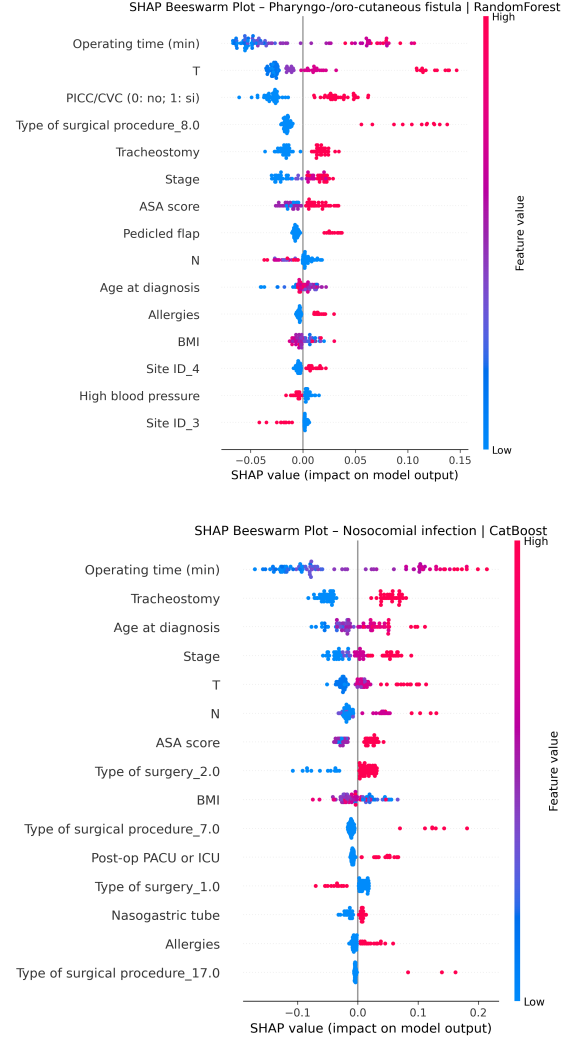


Fig. 3: SHAP beeswarm plots for the selected models on the test set. Top: Pharyngo-/Oro-Cutaneous Fistula (Random Forest); Bottom: Nosocomial Infection (CatBoost). Each point represents an individual patient. The feature “Type of surgery” is encoded as 1 = clean and 2 = clean-contaminated. See Appendix for encoding of SITE ID and Type of Surgical Procedure. Colors indicate normalized feature values (blue = low, red = high). Features are ordered by mean absolute SHAP value.

7 Conclusion

In this study, we investigated the use of machine learning models to predict rare but clinically significant post-operative complications in ENT oncology patients, specifically pharyngo-/oro-cutaneous fistula and nosocomial infections. Our results indicate that ensemble-based models, such as Random Forest and CatBoost, can effectively capture complex interactions among heterogeneous clinical variables, providing reliable support for early risk stratification.

A central contribution of this work is the adoption of a *safety-first* modeling strategy. The framework

prioritizes the early detection of high-risk patients by optimizing decision thresholds based on the positive-class F1-score, which balances recall and precision while still emphasizing sensitivity. This approach ensures that recall remains high—reducing false negatives—while avoiding an excessive number of false positives that could overwhelm clinical resources. This trade-off aligns with clinical practice, where missing a true complication is generally more harmful than raising an additional alert.

Furthermore, the integration of explainability techniques, such as SHAP analysis, enhances transparency by quantifying both the magnitude and direction of feature contributions. This interpretability supports clinician trust and facilitates potential integration of these predictive models into decision-support systems.

7.1 Limitations

Despite the promising findings, several limitations should be considered. First, the target complications are rare events, which makes performance estimates sensitive to small variations in the positive class. Second, the dataset originates from a single tertiary-care center, potentially limiting the generalizability of the results to other hospitals or patient populations with different clinical practices and case mixes.

Additionally, the safety-first operating point, while clinically motivated, increases false positives and may require additional monitoring efforts and resources. Finally, although global interpretability analyses were provided, patient-specific explanations were not fully explored, limiting personalized clinical insights.

7.2 Future Work

Future research will focus on validating the proposed framework on external and temporally distinct cohorts to assess robustness and generalizability. Further investigations will explore probability calibration techniques to produce clinically interpretable risk scores rather than raw prediction probabilities.

From a deployment perspective, future developments include generating per-patient explainability reports that highlight the main risk-driving factors for individual predictions. Finally, the framework could be extended to additional outcomes, such as length of hospital stay or other post-operative complications, supporting broader AI-assisted clinical decision-making in ENT oncology.

References

- [1] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [2] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [3] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

- [4] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [6] T. Saito and M. Rehmsmeier, “The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS ONE*, vol. 10, no. 3, 2015.
- [7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [8] H. Van Dung et al., “Artificial intelligence in surgical training and applications to otolaryngology: A scoping review,” *BMC Medical Informatics and Decision Making*, 2025.
- [9] R. Rajan, A. Smith, et al., “Artificial intelligence in otorhinolaryngology: current trends and application areas,” *European Archives of Oto-Rhino-Laryngology*, 2025.
- [10] J. Park, L. Kim, et al., “Current role of artificial intelligence in head and neck cancer surgery: a systematic review,” *Exploration of Targeted Anti-tumor Therapy*, 2023.
- [11] J. Luu et al., “Practical guide to building machine learning-based clinical prediction models using imbalanced datasets,” *Trauma Surg Acute Care Open*, 2024.
- [12] “Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review,” *Brain Sciences*, 2024.

Appendix

Signed Feature Effect

To improve model transparency and provide clinically meaningful interpretations, we designed a custom signed feature effect analysis. For each feature j , we measured how changing its value affects the model-predicted probability of the positive class, while keeping all other features fixed.

For this, we defined two reference points for each feature: a "low" value (x_j^{low}) and a "high" value (x_j^{high}). For binary features, these correspond to 0 and 1; for continuous features, we used the first and third quartiles (Q1 and Q3) to capture a realistic range of variation and avoid outliers. The signed effect is then computed as the difference in average predicted probabilities between these two points:

$$\Delta_j = \overline{P(y = 1 \mid X_j = x_j^{\text{high}})} - \overline{P(y = 1 \mid X_j = x_j^{\text{low}})}$$

A positive Δ_j indicates that higher values of the feature increase the predicted risk, while a negative Δ_j indicates a decrease.

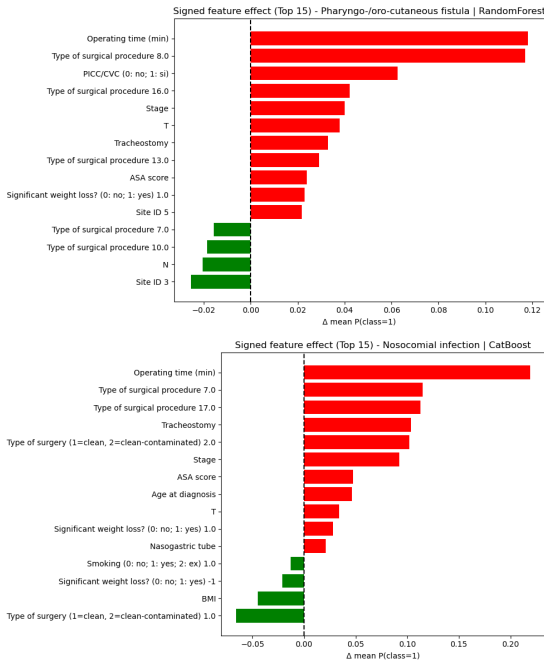


Fig. 4: Signed feature effect for the top predictive features of pharyngo-/oro-cutaneous fistula (top) and nosocomial infection (bottom). Positive values increase predicted risk, negative values decrease it.

Figure 4 reports the signed feature effects for the most predictive variables associated with fistula and nosocomial infection. The plots highlight which features systematically increase or decrease the predicted probability of the positive class across the cohort.

For fistula, features related to surgical complexity and disease severity show the strongest positive effects. Operating time has the largest influence, increasing the

predicted probability by approximately 12% from low to high values. Procedure type 8.0 and the presence of a PICC/CVC also contribute substantially, while TNM stage, T classification, tracheostomy, and ASA score provide additional positive contributions. Conversely, features such as Site ID 3, N classification, and selected procedure types reduce the predicted probability.

For nosocomial infection, operating time again has the largest positive effect, with an estimated increase of approximately 22%. Procedural variables including partial laryngectomy (procedure type 7.0), procedure 17.0, and tracheostomy, together with clinical indicators such as tumor stage, ASA score, and clean-contaminated surgery type, further increase the predicted probability. In contrast, clean surgery type, BMI, and smoking status reduce the predicted risk.

Overall, this custom signed feature effect analysis provides an interpretable, population-level view of how clinical variables influence model predictions, complementing patient-specific explanations such as SHAP and supporting clinically meaningful interpretation of the proposed framework.

Feature Encoding Reference

For transparency and reproducibility, the categorical features used in the SHAP analysis are encoded as follows.

SITE ID: 1: Oral Cavity, 2: Rinofaringe, 3: Oropharynx, 4: Larynx, 5: Hypopharynx, 6: Nose and Paranasal Sinuses, 7: Unknown Focus, 8: Skin, 9: Parotid Gland, 10: Thyroid, 11: Lip, 12: Esophagus, 13: Parapharyngeal Space, 14: Cancered Cyst, 15: Submandibular Gland.

Type of Surgical Procedure: 1: Partial Glossectomy, 2: Hemiglossectomy, 3: Subtotal / Total Glossectomy, 4: Commando, 5: Pharyngectomy, 6: Cordecotomy, 7: Partial Laryngectomy, 8: Total Laryngectomy, 9: Parotid Glandectomy, 10: Emptying only, 11: Thyroidectomy, 12: Rhinectomy, 13: Maxillectomy, 14: Lip Excision, 15: Ear Excision, 16: Pharyngolaryngectomy, 17: Pharyngoglossectomy, 18: Pelvectomy, 19: Other Procedures.